House Price Predictions, Intervals, and Combinations

A Study for Perth in Australia

Zhuoran (Thomas) Zhang

November 16, 2020

University of Aberdeen & Curtin University

Table of Contents

Introduction

What we did?

Automated Valuation Models (AVM), which could provide estimated house price by giving specific variables about properties.

Why?

1. Housing Automated Valuation Service (AVS).
2. Mortgage underwriting, properties refinancing for banking service.
3. Tax assessment; Risk assessment of loan pools.

Introduction

Research Questions

1. Which model is the best or most accurate to provide housing automated valuation service?
2. What is the uncertainty of these predictions from models?
3. Is using the single best model better than models cooperation?

How?

1. Examining models in full samples and rolling windows samples. The Performance of models will be compared.
2. Calculating the predict interval of each observation by quantile measures.
3. Trying the forecast combination (or model averaging).

Literatures

Main Four

1. **Automated Valuation Modelling: A Specification Exercise**. (2013). R. Schulz; M, Wersing and A, Werwatz.
2. **An Introduction to Statistical Learning with Applications in R**. (2013). G, James; D, Witten and T, Hastie.
3. **Semi-parametric Regression with R**. (2018). J, Harezlak; D, Ruppert and M, Wand.
4. **Model Averaging and Its Use in Economics**. (2020). Mark F. J. Steel.

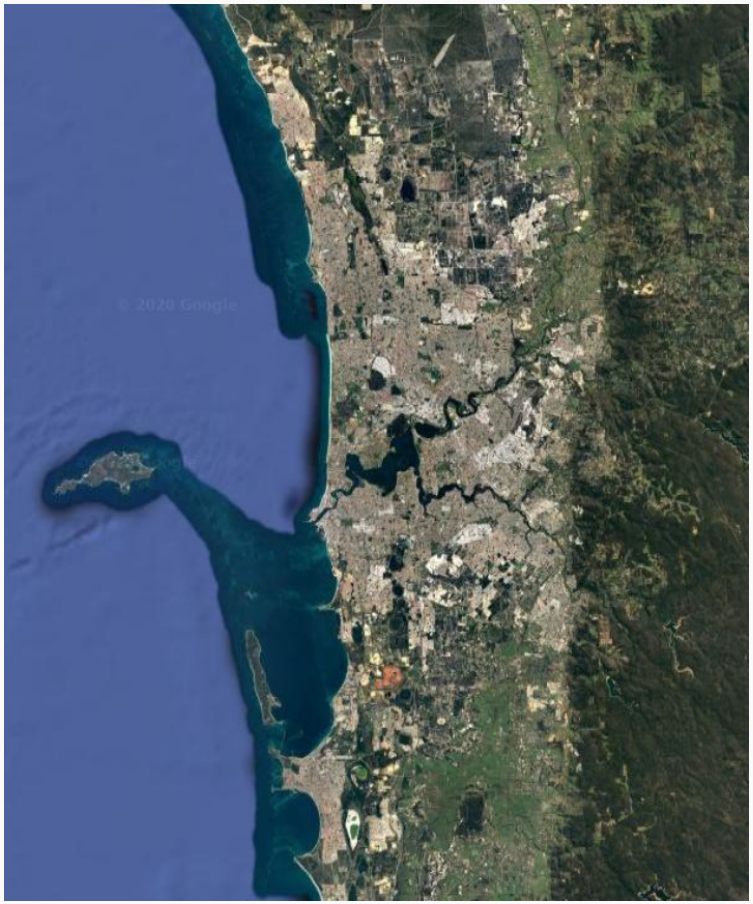
Table of Contents

Real Estate Data in Perth, Australia

Basic Info

- Landgate is our data provider, the Western Australian Land Information Authority.
- Research Period: **1989.01.01 – 2019.12.31**.
- No. of Observations: **755,670** house transactions in total.

Figure 1: The Satellite Shot of Perth



Real Estate Data in Perth, Australia Cont'd

Available Variables

- Transaction price and date.
- Year of build and land area.
- Number of rooms and other facilities.
- Location coordinates.

Data Cleaning

- Restrict between 1 and 6 bedrooms.
- Restrict between 1 and 5 bathrooms.
- Restrict between 0 and 5 the other functional rooms.
- Filter the typos and meaningless observations.

Real Estate Data in Perth, Australia Cont'd

45,558 observations are cleaned. **710,112** observations in final data set.

Table 1: The Summary Statistics of Final Dataset

	Mean	Std.Dev.	Min	Max
Price	379,442	431,851	10,000	50,000,000
Land Area	789.4	560.3	100	10,000
Age	24.5	20.2	0	135
Bedroom	3.3	0.8	1	6
Bathroom	1.5	0.6	1	5
Car Park	1.4	0.8	0	5
Pool	0.15	0.36	0	1

Table of Contents

Models Summary

Parametric Group

1. Basic Linear Model (**Model 1**)
2. Linear Models with Three Estimators (Ridge, LASSO, Elastic-Net; **Model 2-4**)
3. Polynomial Model with Power 10 (**Model 5**)

Semi-parametric Group: Spline Models

1. **Model 6**:

$$ln\ price = Z\beta + f(z_{lon}, z_{lat}) + \epsilon. \tag{1}$$

2. **Model 7**:

$$ln\ price = Z\beta + f(z_{lon}, z_{lat}, z_{landarea}) + \epsilon. \tag{2}$$

Models Summary (Cont'd)

Semi-parametric Group: Spline Models (Cont'd)

3. **Model 8**

$$ln\ price = Z\beta + f(z_{age}) + g(z_{lon}, z_{lat}) + h(z_{landarea}) + \epsilon. \quad (3)$$

4. **Model 9**

$$ln\ price = Z\beta + f(z_{age}) + g(z_{lon}, z_{lat}, z_{landarea}) + \epsilon. \quad (4)$$

Tree Based Group

1. Regression Tree (**Model 10**)
2. Random Forest (**Model 11**), Boosting (**Model 12**)
3. Neutral Network (**Model 13**)

Table of Contents

Table of Contents

Full Sample Analysis

Data Preparation

Training (497,081); Combination (106,515); Testing (106,516).

Tuning Parameters

- Cross Validation (Regularization Estimators, Machine Learning).
- Package *caret* in R.

Full Sample Results

Table 2: The Summary of Full Sample Results

Models	RMSE	MAPE	<1% Relative	<5% Relative
Model 1	0.40120	0.02268	34,350 (32.25%)	96,820 (90.90%)
Model 2	0.40638	0.02305	33,864 (31.79%)	96,343 (90.45%)
Model 3	0.40131	0.02267	34,354 (32.25%)	96,802 (90.88%)
Model 4	0.40234	0.02274	34,237 (32.14%)	96,717 (90.80%)
Model 5	0.32370	0.01735	44,856 (42.11%)	101,458 (95.25%)
Model 6	0.30260	0.01585	48,879 (45.89%)	102,408 (96.14%)
Model 7	0.29673	0.01549	49,833 (46.78%)	102,692 (96.41%)

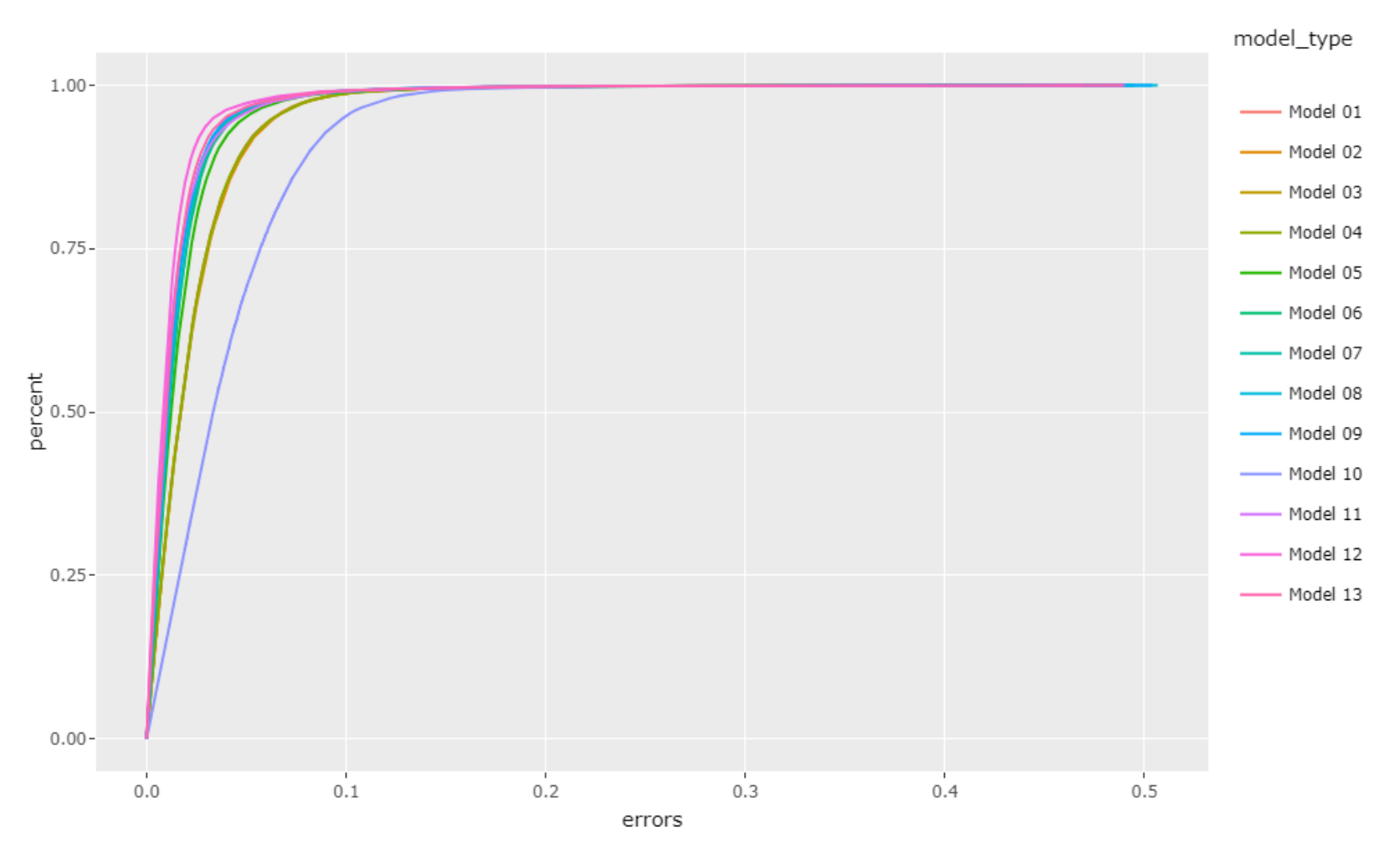
Full Sample Results (Cont'd)

Table 3: The Summary of Full Sample Results, Cont'd

Models	RMSE	MAPE	<1% Relative	<5% Relative
Model 8	0.28991	0.01485	52,838 (49.61%)	102,875 (96.58%)
Model 9	0.28832	0.01473	52,222 (49.03%)	102,788 (96.50%)
Model 10	0.63833	0.04027	16,652 (15.63%)	73,685 (69.18%)
Model 11	0.28931	0.01423	58,348 (54.78%)	102,114 (95.87%)
Model 12	0.25282	0.01211	63,653 (59.76%)	103,607 (97.27%)
Model 13	0.28041	0.01398	55,904 (52.48%)	103,009 (96.71%)

Full Sample Results (Cont'd)

Figure 2: Error Rates of Models



Full Sample Results – Models Comparison

Accuracy:

Parametric Group $<$ Semi-parametric Group $<$ Tree Based Group

Model Interpretability:

Tree Based Group $<$ Parametric Group $<$ Semi-parametric Group

- **Tree Based Group**: Black box process; Predictions.
- **Parametric Group**: Clear and easy process; Parameters, Predictions.
- **Semi-parametric Group**: Clear but complex process; Some parameters, Predictions, **Graphs**.

Full Sample Results – Models Comparison Cont'd

Figure 3: Price Predictor Contours (700 m^2 Land Area)

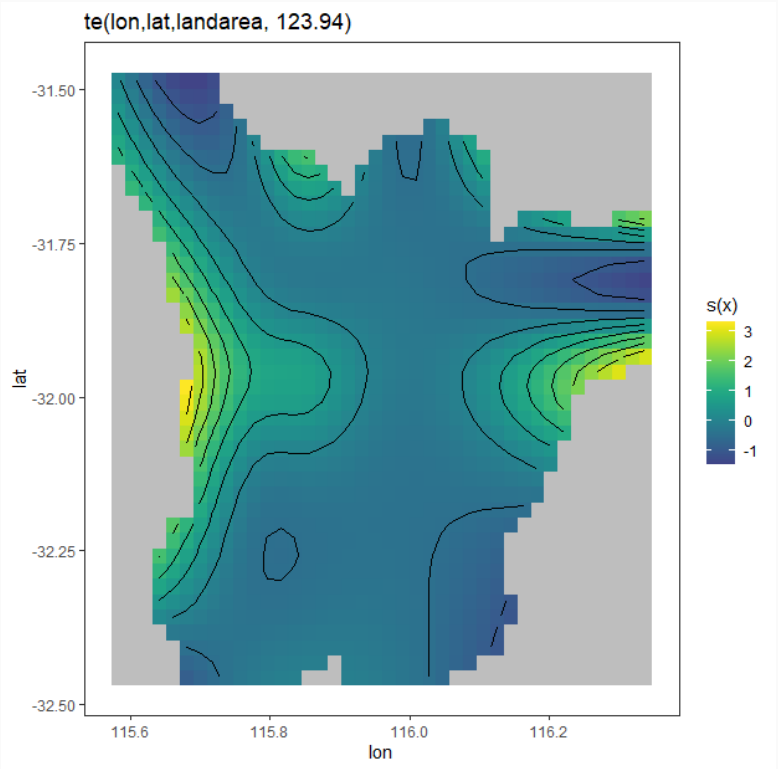
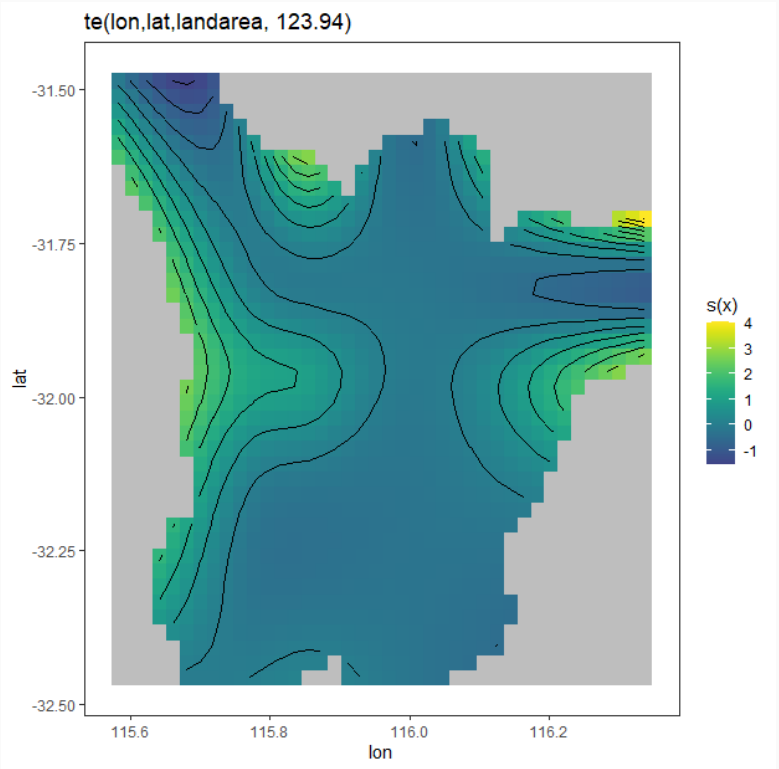


Figure 4: Price Predictor Contours (1000 m^2 Land Area)



Full Sample Results – Prediction Intervals

Table 4: The Summary of Prediction Intervals (95%)

Method	Mean Width	S.D. of Width	Rate within Interval
Linear	0.1191	0.0257	94.97%
Spline	0.0887	0.0222	95.69%
Boosting	0.0730	0.0316	93.91%
Random Forest	0.1230	0.0297	97.26%
Neural Network	0.0808	0.0264	94.76%
Averaging	0.0969	0.0202	97.15%

$$Width = \frac{Upper - Lower}{MarketValue} \quad (5) \qquad Rate = \frac{No.\ fall\ in\ interval}{Total\ No.} \quad (6)$$

Full Sample Results – Prediction Intervals Cont'd

Figure 5: Falling in Interval Versus Width (Random Forest)

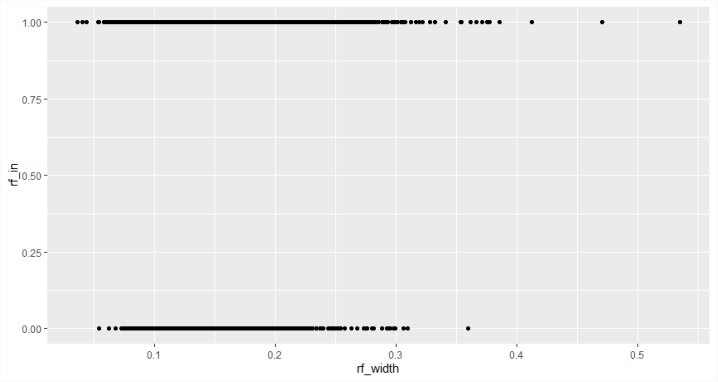


Figure 6: Falling in Interval Versus Width (Boosting)

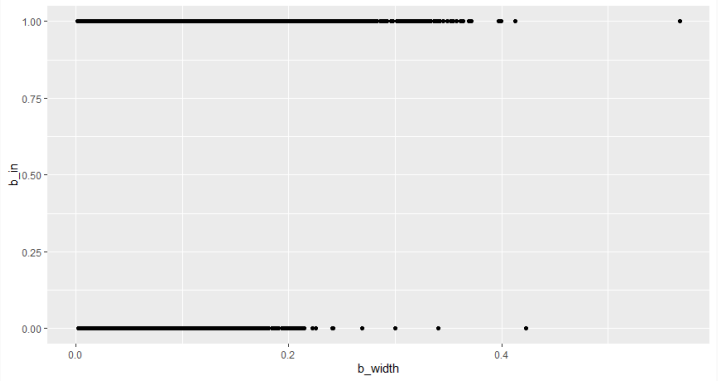
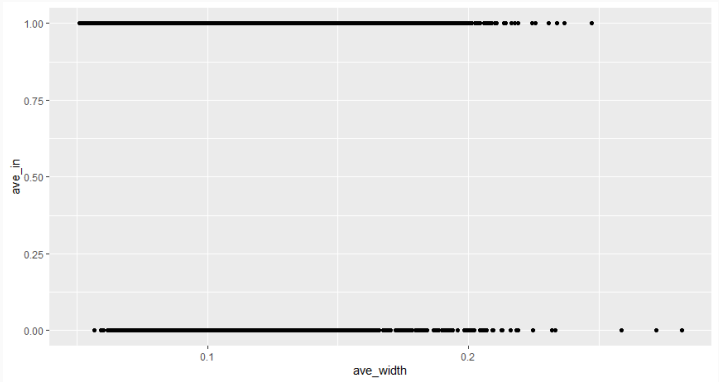


Figure 7: Falling in Interval Versus Width (Averaging)



- When the confidence of prediction intervals are 95%, the predictions should have 5% chance to fall out of the intervals no matter how wide the intervals are.

Full Sample Combination

Combination Methods

Linear Regression Combination; Quantile Regression Combination; Averaging Combination.

Forecasts Selection

All Variables; Best Four Models (9, 11, 12 & 13); LASSO Selection (11, 12 & 13).

Adding Updated Past Sale Price (Extension)

Past prices are updated by local housing price index, then they will be added into combination.

Example:

$$ln\ price = c + (Past\ \ Price) + w_1 forecast_1 + \cdots + w_{13} forecast_{13} + \varepsilon. \quad (7)$$

Full Sample Combination (Cont'd)

Table 5: The Results of Combinations and Their Extension

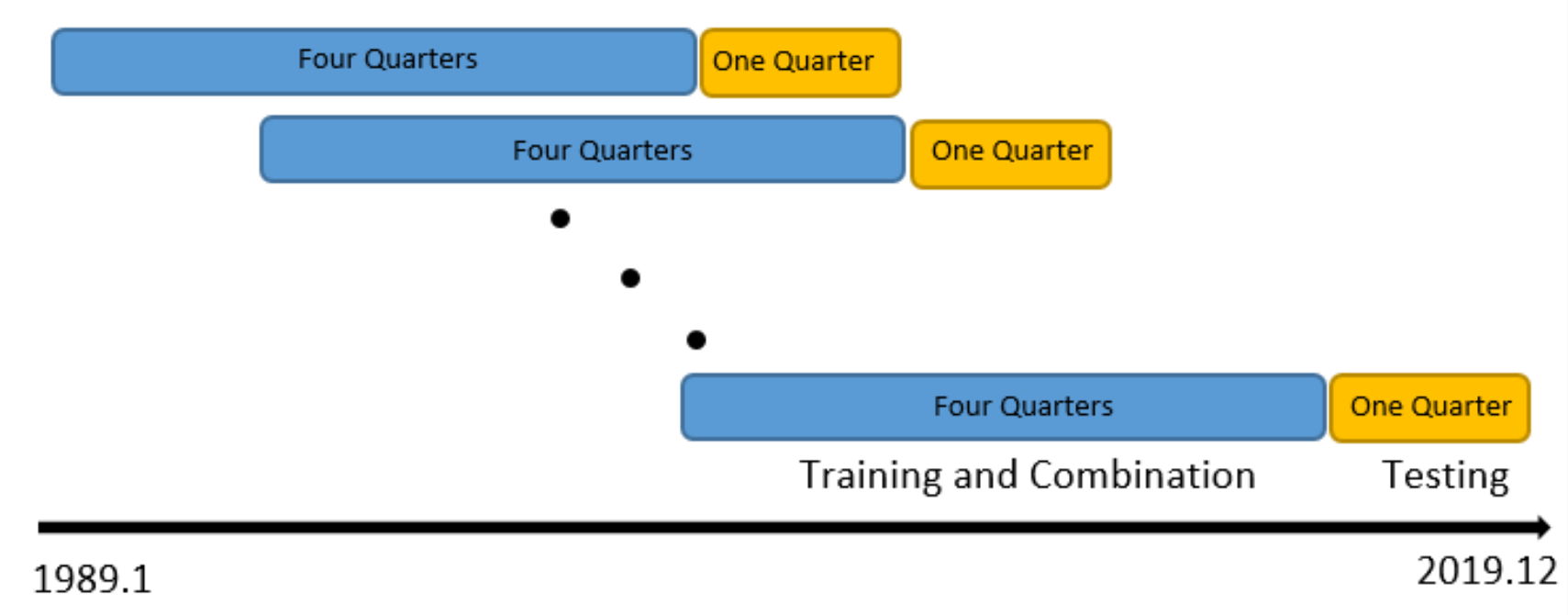
Method	Variables		With Past Sale Prices		Without Past Sale Prices		Overall	
			Extension	Origin	Extension	Origin	Extension	Origin
Linear	All	RMSE	0.23708	0.23907	0.27661	0.27678	0.24909	0.25050
	Forecasts	MAPE	0.01107	0.01122	0.01357	0.13576	0.01179	0.01190
	Best	RMSE	0.23783	0.23984	0.27735	0.27760	0.24984	0.25128
Regression	Four	MAPE	0.01111	0.01126	0.01360	0.01359	0.01182	0.01193
	LASSO	RMSE	0.23791	0.23991	0.27747	0.27769	0.24993	0.25136
	Selection	MAPE	0.01110	0.01126	0.01359	0.01359	0.01182	0.01193
Quantile	All	RMSE	0.23803	0.23993	0.27716	0.27764	0.24991	0.25135
	Forecasts	MAPE	0.01102	0.01115	0.01347	0.01352	0.01172	0.01183
	Best	RMSE	0.23867	0.24054	0.27797	0.27842	0.25061	0.25202
Regression	Four	MAPE	0.01107	0.01120	0.01351	0.01356	0.01177	0.01188
	LASSO	RMSE	0.23865	0.24051	0.27796	0.27840	0.25059	0.25200
	Selection	MAPE	0.01107	0.01120	0.01351	0.01356	0.01177	0.01188
Averaging	All	RMSE	0.28761	0.29381	0.32796	0.32796	0.29977	0.30403
	Forecasts	MAPE	0.01478	0.01518	0.01754	0.01754	0.01558	0.01586
	Best	RMSE	0.26588	0.24914	0.28516	0.28516	0.27157	0.26001
Combination	Four	MAPE	0.01294	0.01176	0.01406	0.01406	0.01327	0.01242
	LASSO	RMSE	0.27774	0.24596	0.28354	0.28354	0.26120	0.25733
	Selection	MAPE	0.01358	0.01156	0.01393	0.01393	0.01259	0.01224

Table of Contents

Rolling Sample Analysis

Data Preparation

Training and Combination Period (Four Quarters; 4:1);
Testing Period (Next Quarter).



In total, there are 120 rolling windows (30 years) examined in this analysis.

Rolling Sample Results

Figure 8: The RMSE of 120 Rolling Windows.

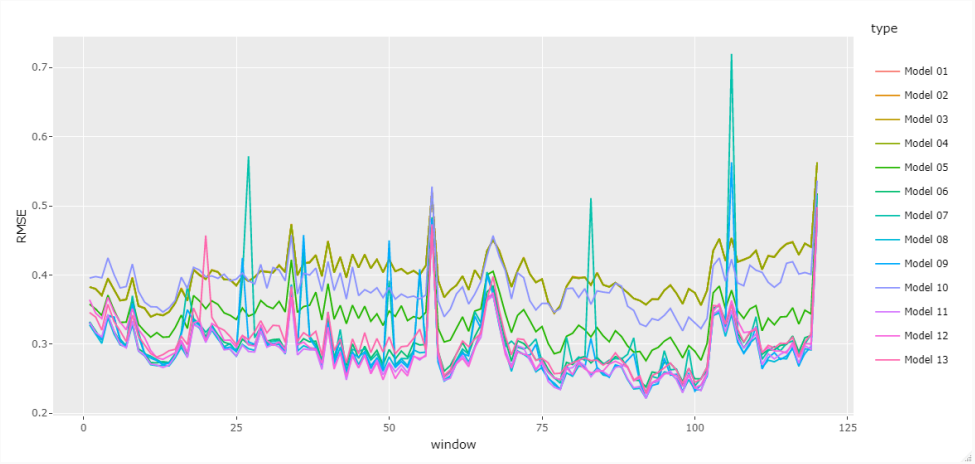
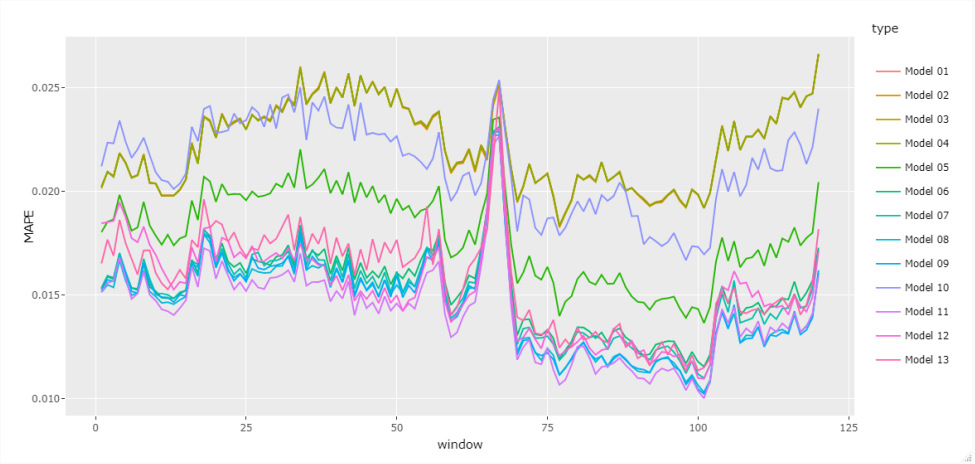


Figure 9: The MAPE of 120 Rolling Windows.



Rolling Sample Results Cont'd

Table 6: Models Pick-up Rates in The 120 Rolling Windows (Full)

Models	Intercept	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Frequency	120	99	49	62	9	55	74
Percent	100%	82.5%	40.83%	51.67%	7.5%	45.83%	61.67%
Models	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
Frequency	65	105	104	98	120	120	77
Percent	54.17%	87.5%	86.67%	81.67%	100%	100%	64.17%

If past price joins the game, to resale observations, the past price is selected in 119 windows; to first sale observations, the LASSO selection combination will rely on Random Forest and Boosting more.

Rolling Sample Results Cont'd

Table 7: Models Pick-up Rates in The 120 Rolling Windows (Past Price)

Models	intercept	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Frequency	120	104	20	54	6	67	55
Percent	100%	86.67%	16.67%	45%	5%	55.83%	45.83%
Models	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
Frequency	57	106	108	82	120	120	85
Percent	47.5%	88.33%	90%	68.33%	100%	100%	70.83%

Table 8: Models Pick-up Rates in The 120 Rolling Windows (First Sale)

Models	intercept	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Frequency	120	59	37	31	5	39	53
Percent	100%	49.17%	30.83%	25.83%	4.17%	32.5%	44.17%
Models	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
Frequency	78	73	84	76	119	120	67
Percent	65%	60.83%	70%	63.33%	99.17%	100%	55.83%

Rolling Sample Results Cont'd

Figure 10: The Combination RMSE of 120 Rolling Windows.

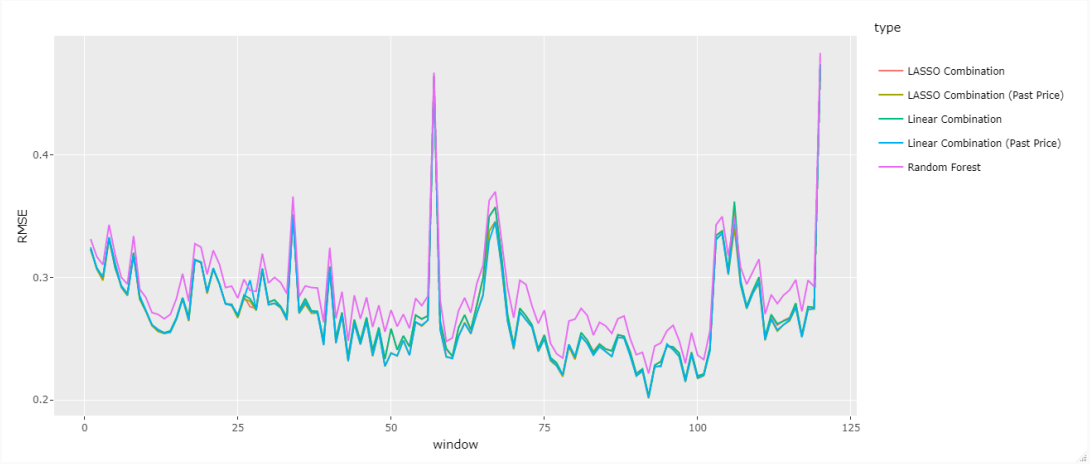


Figure 11: The Combination MAPE of 120 Rolling Windows.

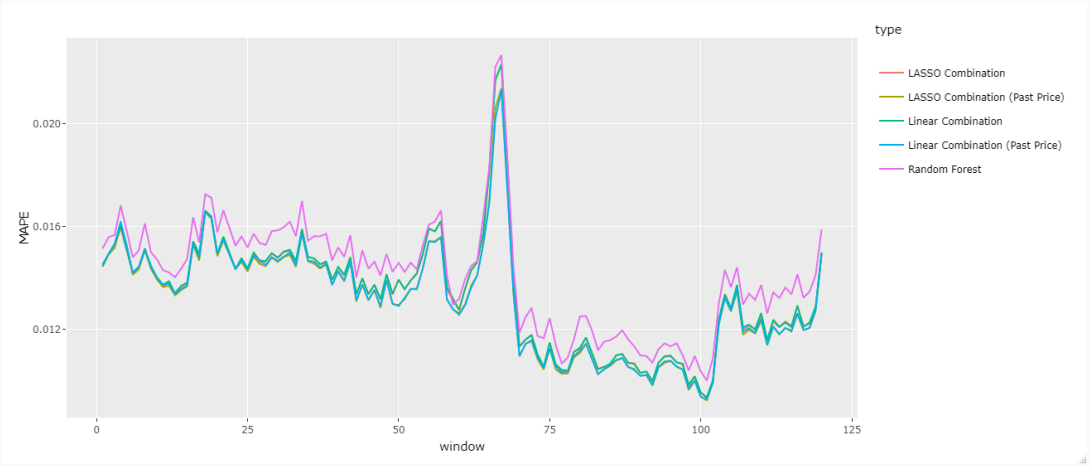


Table of Contents

Conclusions

1. If just considering accuracy, the boosting model is the best. However, if interpretability is also a consideration, the spline model will be the most competitive one without doubts.
2. The combination provides a chance to the cooperation between models and other valuable information. This process can also increase the accuracy of prediction and reduce the uncertainty from models.
3. Some high accuracy models are consistently selected in the 120 rolling windows. Indirectly, it shows that the LASSO selection process works in the forecast combination, it gives similar accuracy comparing with linear combination by less variables. Also, in the rolling windows, it proves that no one can be survived in the huge market changes.

Thank you, any questions or comments?